DOCUMENT RESUME

ED 478 087 IR 058 681

AUTHOR	Bartelt, Bill
TITLE	Analysis and Visualization: Hit or Hype?
PUB DATE	2002-06-00
NOTE	7p.; In: SLA 2002: Putting Knowledge to Work. Papers Presented at the Special Libraries Association Conference (Los Angeles, California, June 9-12, 2002); see IR 058 674.
AVAILABLE FROM	For full text: http://www.sla.org/content/Events/ conference/2002annual/ confpap2002/papers2002conf.cfm/ .
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Data; *Data Analysis; *Information Retrieval; Information Technology; *Visual Aids; World Wide Web
IDENTIFIERS	*Information Value

ABSTRACT

This paper discusses the use of tools for analyzing and visualizing data to synthesize knowledge from data. The first section notes the worsening problem of too much information, resulting from technology advances that have made fact or data look-up fast, efficient, and inexpensive. The second section presents a table of six data types (i.e., structured text, unstructured text, numbers, hierarchical, 2D structures, and 3D structures) and the ways in which corresponding tools enable new value to be extracted. The third section examines content, technology and integration, business and economic, and human issues that affect the success of analysis and visualization tools. (MES)

Analysis and Visualization: Hit or Hype?

Bill Bartelt
Senior Product Manager, CAS
2540 Olentangy River Road, Columbus, Ohio 43202-1505
E-mail: wbartelt@cas.org

Are tools for analysis and visualization a cure for the plague of information overload? If so, why aren't they widely available and widely used? Search and retrieval tools have become very good at finding mountains of information, but tools to deal with the mountain lag behind. The promise of analysis and visualization tools is to aid in navigating, categorizing, summarizing, and finding patterns in the data. Such tools promise us to more quickly extract knowledge and gain insights. What are the challenges we face in turning this promise into reality?

As a producer of and online host to the world's largest databases of chemical, scientific, and technical information, CAS is working to provide tools to help people manage the problem of too much information.

OVERVIEW

Information is everywhere. Information is essential. Information is empowering. And yet, information is elusive. More than ever, we need tools to find and make sense of information around us. Over the past 5-10 years we have seen dramatic improvements in information accessibility, speed, usability, and fact-finding. Expectations have been raised concerning what is possible. However, advancements in the ability to synthesize knowledge from data have not kept pace with the ability to amass information. How have tools for analyzing and visualizing data helped? Content, technology and integration, business and economic, and people issues challenge us in increasing the use of these tools.

SOLVING ONE PROBLEM LEADS TO ANOTHER

It can be said that one long-standing dream has now been satisfied. Technology advances have made fact or data look-up fast, efficient, and relatively inexpensive. The World Wide Web serves as the great integrator of advances in user interface design, data storage, database, server, and networking technology. The proliferation of Web search engines makes it easy for anyone to locate information. Type in a couple of words and the search engine instantly returns hundreds if not thousands of Web pages for perusal. This wonderful capability leads directly to the worsening problem of *too much* information. The ease with which new information is generated, retrieved, and delivered exists for public and private data, unmediated and mediated data. Figures

1 and 2 show the increase in the number of unique Web sites [1], and in the number of abstracts appearing in Chemical Abstracts – over three quarters of a million new abstracts in 2001 alone. [2]

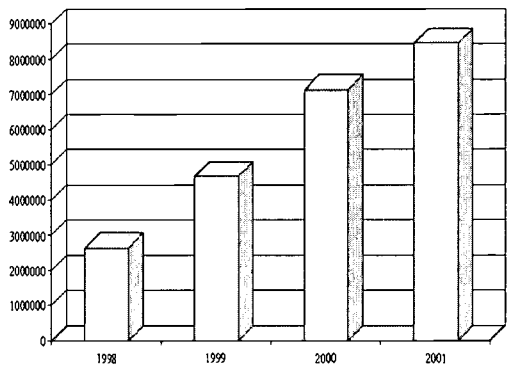

Figure 1 – Unique Web sites per year
Source: OCLC Web Characterization Project

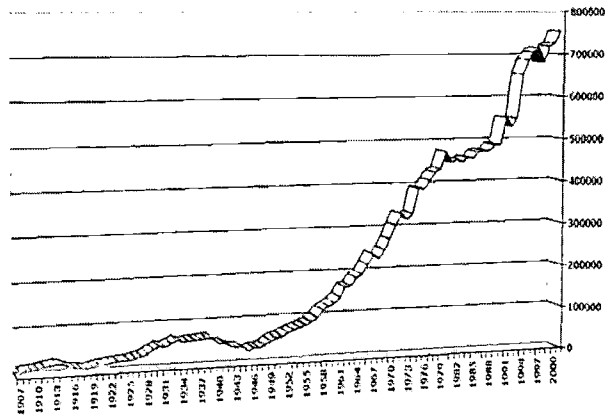

Figure 2 – Abstracts published in CA per year
Source: CAS Statistical Summary 1907-2001

Effective problem solving and decision-making depends on more than the bits of information that are easily found. Information must be gathered, vetted, analyzed, and studied before knowledge, wisdom, and insight are possible. Synthesizing higher order value from raw data takes hard work and time. There is not enough time to read, sort, categorize, analyze, and absorb this much data. New capabilities are needed.

ANALYSIS AND VISUALIZATION TOOLS TO THE RESCUE

Increasingly, tools for analysis and visualization are seen as the solution to this problem. Analysis tools assist in organizing the data, separating it into parts, and studying the interrelationships. Visualization tools present information in ways that reveal the structure, relationships, connections, patterns, and trends in the data. A visualization that is interactive enables rapid navigation through the data, along paths where connections have been made. As the problem of too much information began to emerge, so have new tools for analyzing and visualizing. One way to classify these tools is to consider the types of information they are designed to manage.

Tools tend to be targeted for specific types of data. Examples of data types include structured and unstructured text, numbers, two and three-dimensional chemical substances, hierarchical and network information. Some tools straddle the boundaries of more than one type of data. The table in Figure 3 shows six data types and ways in which corresponding tools enable new value to be extracted. This is not an exhaustive list and new tools and features are routinely announced.

Structured Text: *Organize data, analyze relationships*
- BizInt Smart Charts [3]
- SciFinder Panorama [4]
- STN ANALYZE command [5]
- STN Express with *Discover!* Table Tool [6]
- VantagePoint [7]

Unstructured Text: *Identify and cluster concepts, reveal relationships*
- Aurigin Cartia [13]
- ClearForest ClearResearch [14]
- FAST [15]
- Northern Light [16]
- OmniViz Pro [17]
- Vivisimo [18]

Numbers: *Analyze and reveal relationships, clusters, trends*
- LeadScope [8]
- MS Excel [9]
- Spotfire [10]

Hierarchical: *Reveal relationships, navigate information space*
- Accelrys Diva
- Antarcti.ca Visual.Net [19]
- Inxight Star Tree [20]
- LeadScope
- SmartMoney.com Map of the Market [21]

Structures, 2D: *Visualize chemical models*
- Accelrys Diva [11]
- STN
- STN Easy [12]

Structures, 3D: *Visualize chemical and biological models in three dimensions*
- Accelrys WebLab Viewer [22]
- AutoDOCK [23]
- STN Easy

Figure 3 – Data Types and Related Analysis and Visualization Tools

Among these tools are some of the best available and yet, not all are widely used. Are there tools here you use on a regular basis? Do you rely on analysis and visualization tools to cope with information overload? If not, why not? If there is indeed demand for these capabilities, what will it take for these tools to fulfill that need and achieve widespread use? To find the answers to these questions, we need to look beyond the pretty pictures that first attract our attention.

MANY FACTORS TO SUCCESS

Analysis and visualization features and functions vary greatly, but so do several other key factors not immediately considered. The future success of analysis and visualization tools lies in vendors' ability to meet customer's needs in achieving the proper balance of content, technology and integration, business and economic issues. Not to be overlooked are the human factors. In fact, there is no single issue that guarantees success but there are many issues that can hinder it. The relative importance of each issue will vary according to the needs of the individual organization's situation. Let's examine the issues in more detail.

1. Content

The source of the information to be analyzed and visualized is at the center of the problem and critical to any solution. If the source of data is private, then it may be presumed that the data will be hosted in-house. The data needs to be accessible to and in a format expected by the tool. To enable this may require a certain level of computer system and network expertise. If the information is from a public source, it should be as comprehensive, consistent, and of high quality as possible. In order to demonstrate their capabilities, many vendors apply their tools to publicly available information, but also support in-house implementations for private data.

Public Sources:

- Web pages gleaned from public Internet sites
- Government-produced files such as Medline and the U.S. Patent and Trademark Office database,
- The Open Directory Project Web index
- Stock market data

Vendor Databases:

- Chemical Abstracts Service databases such as CAplus, REGISTRY, and CASREACT
- Value-added patent databases from Derwent
- Business, industry, and news databases

In the operation of its services, CAS supplies information for its tools in the form of databases it builds from publicly disclosed journal and patent sources.

2. Technology and Integration

Hardware and software architectures affect the speed and scalability of analysis and visualization. When response time needs and data set sizes exceed desktop capacities, analysis may need to be handled remotely in a multi-tiered architecture. This in turn may require a substantial investment in computer hardware and networking. Data may have to be accessed, parsed, cleaned, and reformatted before it can be analyzed. The visualization is typically delivered to the individual's desktop. More complicated solutions may be needed when data is from multiple sources, especially if those sources are a combination of internal and external data.

In the information environment, it is essential that the analysis and visualization tools be integrated with the search and retrieval tools. It should be seamless to invoke the necessary tools for the job. Any extra hoops that must be jumped through are barriers to success. With SciFinder, CAS has made a variety of tools available to seamlessly analyze and visualize search results. At times, the integration extends to third-party tools such as Microsoft Excel, Aurigin Aureka, and Spotfire DecisionSite.

3. Business and Economics

The cost to enable analysis and visualization can be daunting. In this emerging area, advanced software tends to be expensive. Besides the added hardware, networking and systems development costs already mentioned, the cost of data acquisition must be considered. CAS has innovated new solutions to help its customers deal with data acquisition costs. The STN ANALYZE command utilizes a tiered pricing structure which caps the data expense after the first 10,000 database records. With SciFinder task and subscription pricing, analysis of search results is supported at no additional cost.

Finally, the ongoing support and maintenance costs should not be overlooked. Because the benefits are sometimes difficult to quantify, the return on investment may be difficult to calculate.

Because this is an emerging field, there are few established players. An important business consideration may be in the vendor's ability to provide systems integration, training and support. In the aftermath of the dot-com collapse, there is added awareness of the stability and long-term economic viability of software vendors.

4. The Human Factor

An important factor in fostering the use of analysis and visualization tools is training. The sooner those who need the tools can make effective use of them, the sooner a return on the investment can begin. Power tools that are complicated or difficult-to-understand are usually in the realm of a small group of information specialists. However, when tools are needed by a larger audience, those that are easy to use have lower training costs and will gain quicker acceptance. In considering this, the software interface should be easy to use and the resulting analysis or visualization should be easy to interpret. Without a doubt, STN is an information specialist's tool, but frequent training seminars, newsletters, and the best Help Desk in the business help customers stay up to date with the necessary skills. SciFinder is well known for its ability to be used with little or no training

SUMMARY

As problems in information retrieval have been increasingly solved over the past several years, the complications of information overload have come to the forefront. Many innovative and powerful solutions have emerged in the form of tools for analysis and visualization of information. The future success of these tools is dependent on much more than their exciting and effective ways of presenting information. A proper balance of content, technology and integration, business and economic issues is needed.

Endnotes:

[1] http://wcp.oclc.org/

[2] http://www.cas.org/EO/casstats.pdf
[3] http://www.bizcharts.com/
[4] http://www.cas.org/SCIFINDER/panorama.html
[5] http://www.cas.org/ONLINE/STN/STNOTES/stnote17.html
[6] http://www.cas.org/ONLINE/STN/discover.html
[7] http://www.thevantagepoint.com
[8] http://www.leadscope.com
[9] http://www.microsoft.com/office/excel/default.asp
[10] http://www.spotfire.com
[11] http://www.accelrys.com/products/diva/index.html
[12] http://www.cas.org/stn.html
[13] http://www.aurigin.com
[14] http://www.clearforest.com
[15] http://www.alltheweb.com
[16] http://www.northernlight.com
[17] http://www.omniviz.com
[18] http://vivisimo.com/
[19] http://antarcti.ca
[20] http://www.inxight.com/
[21] http://www.smartmoney.com/marketmap/
[22] http://www.accelrys.com/viewer/index.html
[23] http://www.scripps.edu/pub/olson-web/doc/autodock/

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

X	This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.
	This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").